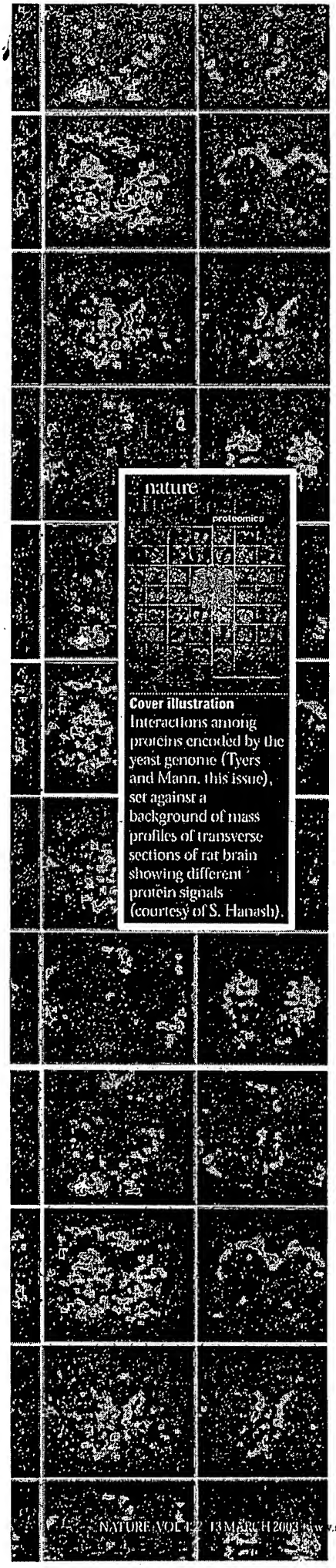


Cover illustration
Interactions among proteins encoded by the yeast genome (Tyers and Mann, this issue), set against a background of mass profiles of transverse sections of rat brain showing different protein signals (courtesy of S. Hanash).

nature insight proteomics

We are only just beginning to appreciate the power and limitations of the genomics revolution, yet hard on its heels proteomics promises an even more radical transformation of biological and medical research. Encoded proteins carry out most biological functions, and to understand how cells work, one must study what proteins are present, how they interact with each other and what they do.

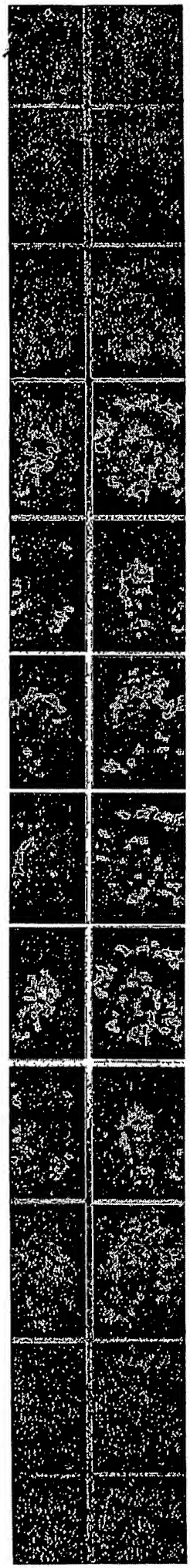
The term proteome defines the entire protein complement in a given cell, tissue or organism. In its wider sense, proteomics research also assesses protein activities, modifications and localization, and interactions of proteins in complexes. It is very much a technology-driven enterprise, and this collection of reviews reflects the progress made and future developments needed to identify proteins and protein complexes in biological samples comprehensively and quantitatively with both high sensitivity and fidelity.

By studying global patterns of protein content and activity and how these change during development or in response to disease, proteomics research is poised to boost our understanding of systems-level cellular behaviour. Clinical research also hopes to benefit from proteomics by both the identification of new drug targets and the development of new diagnostic markers.

Like genomics, the sheer scale of proteomics research makes it a community effort with the Human Proteome Organisation (HUPO) playing an important role in coordinating proteomics projects worldwide. The wealth of information produced poses challenges for data management, and necessitates publicly accessible databases that use agreed standards to describe protein data, allowing data comparison and integration. Furthermore, the expense and scale of proteomics technologies restricts their access, and solutions must be found that allow the widespread use of proteomics tools. In this spirit, in a commentary published in today's issue of *Nature* (422, 115–116; 2003), Ruedi Aebersold proposes a community-wide strategy that could help shift proteomics research towards a 'browsing mode' of searching through existing information.

We are pleased to acknowledge the financial support of Amersham Biosciences in producing this Insight. As always, *Nature* carries sole responsibility for the editorial content and peer review.

Barbara Marte Senior Editor



From genomics to proteomics

Mike Tyers* & Matthias Mann†

**Samuel Lunenfeld Research Institute, Mount Sinai Hospital, and Department of Medical Genetics and Microbiology, University of Toronto, Toronto, Canada M5G 1X5 (e-mail: tyers@mshri.on.ca)*

†Center for Experimental BioInformatics, Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark (e-mail: mann@bmb.sdu.dk)

Proteomics is the study of the function of all expressed proteins. Tremendous progress has been made in the past few years in generating large-scale data sets for protein–protein interactions, organelle composition, protein activity patterns and protein profiles in cancer patients. But further technological improvements, organization of international proteomics projects and open access to results are needed for proteomics to fulfil its potential.

The term proteome was first coined to describe the set of proteins encoded by the genome[1]. The study of the proteome, called proteomics, now evokes not only all the proteins in any given cell, but also the set of all protein isoforms and modifications, the interactions between them, the structural description of proteins and their higher-order complexes, and for that matter almost everything 'post-genomic'. In this overview we will use proteomics in an overall sense to mean protein biochemistry on an unprecedented, high-throughput scale. The hope, now being realized, is that this high-throughput biochemistry will contribute at a direct level to a full description of cellular function.

Proteomics complements other functional genomics approaches, including microarray-based expression profiles[2], systematic phenotypic profiles at the cell and organism level[3,4], systematic genetics[5,6] and small-molecule-based arrays[7] (Fig. 1). Integration of these data sets through bioinformatics will yield a comprehensive database of gene function that will serve as a powerful reference of protein properties and functions, and a useful tool for the individual researcher to both build and test hypotheses. Moreover, large-scale data sets will be crucial for the emerging field of systems biology[8].

Challenges and approaches in proteomics

Proteomics would not be possible without the previous achievements of genomics, which provided the 'blueprint' of possible gene products that are the focal point of proteomics studies. Although almost trite, the tasks of proteomics can usefully be contrasted with the huge but straightforward challenges initially facing the genome projects. Unlike the scalable exercise of DNA sequencing, with its attendant enabling technologies such as the polymerase chain reaction and automated sequencing, proteomics must deal with unavoidable problems of limited and variable sample material, sample degradation, vast dynamic range (more than 10^6-fold for protein abundance alone), a plethora of post-translational modifications, almost boundless tissue, developmental and temporal specificity, and disease and drug perturbations. While proteomics is by definition expected to yield direct biological insights, all of these difficulties render any comprehensive proteomics project an inherently intimidating and often humbling exercise.

In this *Nature* Insight, five central pillars of proteomics research are discussed with an emphasis on technological developments and applications. These areas are mass spectrometry-based proteomics, proteome-wide biochemical assays, systematic structural biology and imaging techniques, proteome informatics, and clinical applications of proteomics. As is apparent from the reviews, the divisions between these areas are somewhat arbitrary, not least because technological breakthroughs often find immediate application on several fronts. More important, biologically useful insights into protein function often emerge from the combination of different proteomic approaches.

Mass spectrometry-based proteomics

The ability of mass spectrometry to identify ever smaller amounts of protein from increasingly complex mixtures is a primary driving force in proteomics, as described in the review on page 198 by Aebersold and Mann. Initial proteomics efforts relied on protein separation by two-dimensional gel electrophoresis, with subsequent mass spectrometric identification of protein spots. An inherent limitation of this approach is the depth of coverage, which is necessarily constrained to the most abundant proteins in the sample. The rapid developments in mass spectrometry have shifted the balance to direct mass spectrometric analysis, and further developments will increase sensitivity, robustness and data handling.

The past year has seen partial analysis of the yeast interactome, the malaria proteome, bacterial proteomes and various organellar proteomes (see review by Aebersold and Mann, page 198). These vast data sets represent but the tip of the iceberg for biological discovery and drug development. An enormous challenge resides in the obvious fact that the proteome is a dynamic, not a static, entity. Initial efforts to gauge proteome-wide regulatory events in single experiments have been directed at the yeast phosphoproteome[9] and the ubiquitin-mediated 'degradome' (S. P. Gygi, personal communication). Much higher throughput and sensitivity will be needed to enable true proteome dynamics and moment-by-moment snap shots of cellular responses. Nascent methods for gel-free analysis of complex mixtures hold great promise in this regard[10]. Further needs will include more complete sequence coverage of each individual protein, robust and varied methods for sample preparation, and sophisticated algorithms for automated protein identification and detection of post-translational modifications. The ambitious goals of systems biology, which aims to comprehensively model cellular behaviour at the whole-system level[8,11], will also require reliable quantitative methods.

Array-based proteomics

A number of established and emergent proteome-wide platforms complement mass spectrometric methods, as

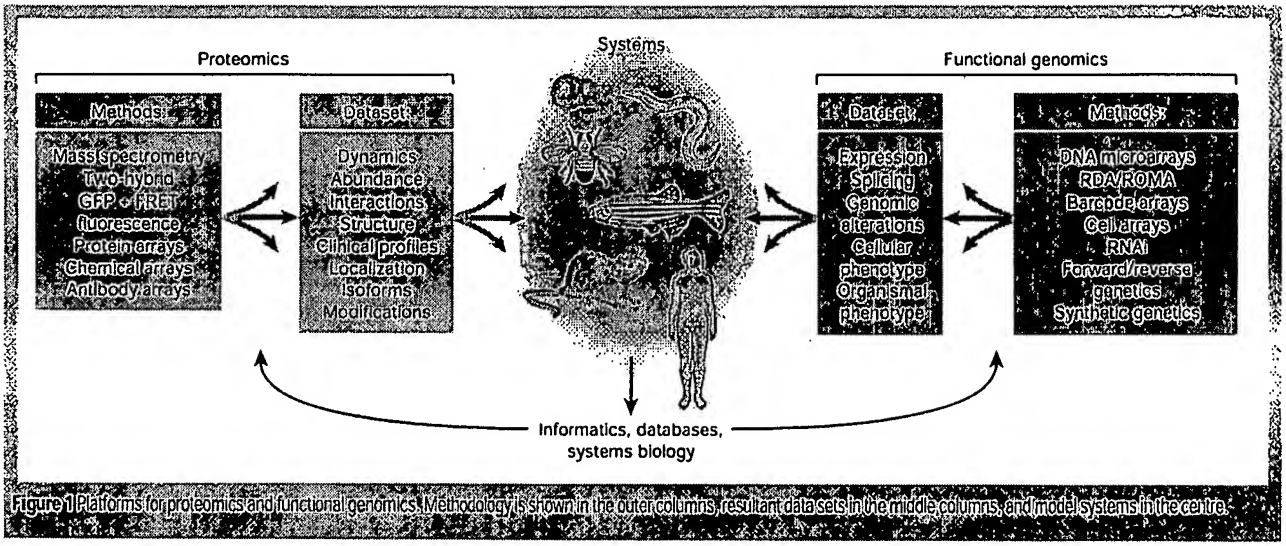


Figure 1 Platforms for proteomics and functional genomics. Methodology is shown in the outer columns, resultant data sets in the middle columns, and model systems in the centre.

reviewed on page 208 of this issue by Stan Fields and co-workers. The forerunner amongst these efforts is the systematic two-hybrid screen developed by Fields[12]. Unlike direct biochemical methods that are constrained by protein abundance, two-hybrid methods can often detect weak interactions between low-abundance proteins, albeit at the expense of false positives.

More recently, various protein-array formats promise to allow rapid interrogation of protein activity on a proteomic scale. These arrays may be based on either recombinant proteins or, conversely, reagents that interact specifically with proteins, including antibodies, peptides and small molecules[13]. Readouts for protein-based arrays can derive from protein interactions, protein modifications or enzymatic activities. A current challenge is to effectively couple high-end mass spectrometry to array formats. Array-based approaches can also use *in vivo* readouts, for example in the systematic analysis of protein localization in the cell through green fluorescent protein (GFP) signals or protein association through fluorescence resonance energy transfer (FRET) between protein fusions to different wavelength variants of GFP. Finally, cell- and tissue-based arrays enable yet another layer of functional interrogation.

One practical bottleneck to these approaches, and indeed to most systematic approaches, has been the limited availability of validated genome-wide complementary DNA for use in the capture of protein complexes with epitope tags. The FlexGene consortium between academic institutions and industry aims to develop complete cDNA collections in recombination-based cloning formats for the biomedical community (see http://www.hip.harvard.edu).

Structural proteomics

Beyond a description of protein primary structure, abundance and activities, the ambitious goal of systematically understanding the structural basis for protein interactions and function is reviewed by Baumeister *et al.* on page 216 of this issue. Through literary metaphor, the authors make a compelling argument that a full description of cell behaviour necessitates structural information at the level not only of all single proteins, but of all salient protein complexes and the organization of such complexes at a cellular scale. This all-encompassing structural endeavour spans several orders of magnitude in measurement scale and requires a battery of structural techniques, from X-ray crystallography and nuclear magnetic resonance (NMR) at the protein level, to electron microscopy of mega-complexes and electron tomography for high-resolution visualization of the entire cellular milieu. The recurrent proteomic theme of throughput and sensitivity runs through each of these structural methods, and Baumeister *et al.* suggest novel solutions, even including eliminating the crystals from crystallography! NMR and *in silico* docking will be necessary to build in dynamics of protein interactions, much of which may be controlled through largely unstructured regions[14].

Informatics

As with any data-rich enterprise, informatics issues loom large on several proteomics fronts. On page 233 of this issue, Boguski and McIntosh highlight the importance of sample documentation, the implementation of rigorous standards and proper annotation of gene function[15]. It is crucial that software development is linked at an early stage through agreed documentation, XML-based definitions and controlled vocabularies that allow different tools to exchange primary data sets. Considerable effort has already gone into interaction databases[16] and systems biology software infrastructure[17] that should be built upon by future proteomics initiatives. The development of statistically sound methods for assignment of protein identity from incomplete mass spectral data will be critical for automated deposition into databases, which is currently a painstaking manual and error-prone process. Lessons learned from analysis of DNA microarray data, including clustering, compendium and pattern-matching approaches, should be transportable to proteomic analysis[2], and it is encouraging that the European Bioinformatics Institute and the Human Proteome Organisation (HUPO) have together started an initiative on the exchange of protein–protein interaction and other proteomic data (see http://psidev.sourceforge.net/)

Clinical proteomics

Proteomics is set to have a profound impact on clinical diagnosis and drug discovery, as is fittingly reviewed by Sam Hanash on page 226, the inaugural president of HUPO. Because most drug targets are proteins, it is inescapable that proteomics will enable drug discovery, development and clinical practice. The form(s) in which proteomics will best fulfil this mandate is in a state of flux owing to a multitude of factors, not the least of which are the varied technological platforms in different stages of implementation.

The detection of protein profiles associated with disease states dates back to the very beginning of proteomics, when two-dimensional gel electrophoresis was first applied to clinical material. The advent of mass spectrometers now able to resolve many tens of thousands of protein and peptide species in body fluids is set to revolutionize protein-based diagnostics, as demonstrated in recent retrospective studies of cancer patients[18]. The robust and high-throughput nature of mass spectrometric instrumentation is imminently suited to clinical applications. Protein- and antibody-based arrays with validated diagnostic readouts may also become amenable to the clinical